DOCUMENT RESUME

ED 227 175 TM 830 204

AUTHOR Choppin, Bruce H.
TITLE Extracting More Information from Multiple Choice Tests: Analytic Techniques for the Answer-until-Correct Mode.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Apr 83
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, Canada, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Assisted Testing; Difficulty Level; *Guessing (Tests); Knowledge Level; *Latent Trait Theory; *Mathematical Models; *Measurement Techniques; *Multiple Choice Tests; Psychometrics; Test Items
IDENTIFIERS *Answer Until Correct; Distractors (Tests); One Parameter Model; *Partial Knowledge (Tests); Rasch Model

ABSTRACT

In the answer-until-correct mode of multiple-choice testing, respondents are directed to continue choosing among the alternatives to each item until they find the correct response. There is no consensus as to how to convert the resulting pattern of responses into a measure because of two conflicting models of item response behavior. The first suggests that partial knowledge allows the subject to eliminate some distractors immediately, and then assumes essentially random guessing among the remainder. The second proposes that the first error made by the subject results from misinformation, but that guessing comes into play after that. The paper considers three latent trait measurement models from each of these perspectives. Each is an extension of the Rasch one-parameter logistic model. The first, which is most relevant to the partial knowledge viewpoint, is based on a count of the error choices before the correct response is identified. The second calibrates the difficulty of each step in each item. The third calibrates the difficulty of each distractor. It is argued that the second model provides the best context for distinguishing between the misinformation and partial knowledge approaches. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227175

EXTRACTING MORE INFORMATION FROM MULTIPLE CHOICE TESTS: ANALYTIC TECHNIQUES FOR THE ANSWER-UNTIL-CORRECT MODE

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

by

Bruce H. Choppin
Center for the Study of Evaluation
University of California, Los Angeles

Paper read at the Annual Meeting of the
American Educational Research Association,
Montreal, April 1983

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B H. Choppin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TM 830 204

EXTRACTING MORE INFORMATION FROM MULTIPLE CHOICE TESTS: ANALYTIC TECHNIQUES FOR THE ANSWER-UNTIL-CORRECT MODE

1. Introduction

Though they are convenient to use and have some desirable psychometric properties, multiple choice tests have been widely attacked (Wood, 1977). Three specific criticisms that have been made against conventional multiple choice tests are:

1) That they face the testee with three or four times as many incorrect statements as correct ones and provide no feedback to help the student learn the correct answers.

2) That they encourage random guessing.

3) That they are inefficient in that little information is gained about the student from his response to a single item.

The "answer-until-correct" testing mode (Brown, 1965; Hanna, 1975) is designed to overcome these problems. In this mode the student is presented with instant feedback to a response. If the response is correct, the student is directed to continue to the next question, but if the response is incorrect he or she is asked to attempt the item again. This form of testing has the advantage of extracting significantly more information about a student's ability from a given number of items, and thus makes it easier to distinguish between different levels of partial knowledge or part mastery. It has also been suggested that this response mode may reduce the incidence of random guessing behavior among students, and it has the additional benefit that (most of the time) the final answer chosen by the student

to an item is also the correct one. There is, a priori, reason to believe that this response, the one that receives positive reinforcement, is the one most likely to be remembered.

A number of research studies have focused on the characteristics and usefulness of answer-until-correct testing. For example, Merwin (1959), Brown (1965) and Frary (1980) investigated various scoring procedures. None of the more complex alternatives they tried appeared to improve significantly on Brown's simple approach of reducing the total score by one point for every incorrect distractor selected. Hanna (1975), and Kane & Moloney (1978), investigated the implications of AUC responding for reliability and validity. Hanna suggested that the AUC procedure increased reliability but generally appeared to decrease validity (as measured by correlation with a substantive external criterion). The implication is that testwiseness may play a more significant role on AUC tests than on conventional tests. This relates back to Merwin's earlier paper in which he concluded that if test constructors were to reap the potential advantages of the AUC procedure, then item distractors would have to be carefully designed so as to relate in a clear way to the criterion variable.

Much of the earlier work reported on this topic displayed considerable vagueness as to the presumed behavior of the student when taking a test.

A careful reading and analysis of the logic presented suggests that the writers were assuming the relevance of one or the other of two contrasting and incompatible models. The first, which may be

called the partial knowledge model, assumes that the student may know enough about the subject matter with which the item is concerned in order to be able to eliminate one or more of the distractors with some certainty. He is then presumed to guess at random among those that remain. Complete mastery of the problem involves the certain elimination of all but one of the alternative responses so that the student chooses the correct answer without guessing.

The second model assumes that a student arrives at an incorrect response not through some guessing procedure, but through the application of misinformation. Under the answer-until-correct procedure, such a student having applied his misinformation to obtain the wrong answer, is forced to choose again. The feedback that the first piece of misinformation is incorrect may provide important incidental learning. The next choice may be a random guess, or another response selected on the basis of misinformation.

Frary showed that the AUC procedure was effective in discriminating between students when they operated on the basis of partial information, but suggested that the scoring procedure could be improved for students operating the misinformation model. Wilcox (1982) further considers the distinction between the partial knowledge and misinformation models and discusses appropriate rules for scoring tests when the latter operates. Unfortunately, it would appear that in practice many individuals use both strategies when taking tests, and it is difficult to tell when looking at the pattern of results on which items they were employing partial knowledge and on which

misinformation. Questioning students following the administration of an AUC test could help to clarify this issue.

The answer-until-correct procedure has made comparatively little impact on the field of educational testing in the seventeen years since Brown's paper for two reasons:

(a) the lack of convenient and appropriate technology for providing instant feedback to the student, since clinical administration of tests is prohibitively expensive; and

(b) the absence of a sound theoretical base for turning the data into measures, for while Brown's system appears to work in practice, there is no model to substantiate it or check its validity.

On the first issue, there have been a number of recent developments. Answer-until-correct tests currently in use (on an experimental or regular basis) use one of three different feedback technologies. The first approach requires an answer sheet preprinted in invisible ink, so that when the student responds (using a special pen) a portion of the preprinted material becomes visible, and the student obtains the appropriate feedback. The second method involves having the student erase a shield printed over the top of the feedback information again on a specially prepared answer sheet. Each of these approaches requires some special equipment for preparing the answer sheets which have to be customized to fit a particular test. However, this equipment is now fairly generally available, and the answer sheets produced from it are not unduly expensive.

The third approach involves testing by the computer. This method is potentially superior to the other methods because it allows the

recording of the sequence in which particular responses are chosen. The first two methods described permit the inference that the correct response was chosen last, but do not easily allow the earlier incorrect responses to be ordered. Until very recently the computer was far too expensive to be considered seriously for routine use as a test administering device, but the rapid development of terminals and in particular of inexpensive micro processers opens up new possibilities.

The computer is able not only to record the sequence in which distractors are selected, but also to accumulate other information (e.g., how long was the delay between each response), and continually update estimates of the student's level of performance and the measurement precision. It is also able to provide more or less detailed feedback under the control of the test constructor, and to provide the feedback in an entirely standard fashion so that no inadvertant clues are presented. During the last year, a team at the Center for the Study of Evaluation has devoted considerable effort to developing an effective and efficient program for administering answer-until-correct tests using Apple microcomputer systems. We have designed this system so as to be useful to classroom teachers who currently have access to Apple or similar computers, and also to us in collecting answer-until-correct data for our psychometric research.

The rest of this paper will be devoted to describing the latent trait models which address the second of the problems mentioned

earlier, the absence of a sound theoretical base for turning the response data into a measure.

2. Latent Trait Models

Three new latent trait models will be described. They differ from one another in their complexity, though each is designed to yield a single parameter to measure student achievement.

The simplest, a "partial credit" model has a single difficulty parameter for each item. It is the latent trait analogue for Brown's (1965) integer scoring scheme based on the number of attempts needed to reach the correct response. The scoring is from 1.0 for a correct response on the first attempt to 0.0 for failure in (m-1) attempts, where there are m alternatives presented for an item (see Figure 1). This model takes no account of the variations in distractor attractiveness from item to item, nor of which distractors were actually selected by the respondent.

The second latent trait model treats the test as a sequence of distinct steps each of which has a difficulty parameter. A single five-way multiple choice item can be regarded as comprising four steps, with each successive step after the first being attempted if, and only if, the preceding one is failed. The scoring is 1/0 for each step, with steps not attempted being coded as incomplete data (Figure 2). This produces four difficulty parameters for each item, but a single and more precise ability estimate for the individual. The method does not assume that all the items have the same logical

structure with regard to difficulty, but it takes no account of exactly which distractors are selected.

The third model is an extension of the second. In this model, the step difficulty values for an item vary in terms of which distractors were previously selected. Thus for a five-way multiple choice item there is one difficulty parameter at the first step, four at the second, six at the third, four at the fourth. This give a total of fifteen difficulty parameters for a single five-way multiple choice item. It should in general give a better fit than the model described above because it treats the distractors individually, but it requires more data for the necessary calibration of the item parameters.

To some extent, the utility of these models is going to depend on the relative preponderance of the two styles of student behavior discussed earlier. Under partial knowledge, distractor elimination and random guessing (style A) the noise introduced by guessing precludes the possibility of very precise measurement, and the first model described may well prove as effective as either of the others. Where item responses based on correct information or misinformation (style B) dominate, we would expect that models two and three would provide more precise and valid measures of student performance.

Each of the models described is based on the simple one-parameter Rasch logistic model. This is for two reasons. Firstly, the Rasch model seems the logical choice in a situation which involves the construction of new test instruments, since it focuses attention on meeting the logical rquirements for objective measurement. Secondly,

the main alternative, the three-parameter logistic model, has severe practical limitations even when applied to regular test data. Estimating techniques are primitive, and very large samples are required in order to obtain stable parameter estimates. The three-parameter model has been found useful in describing large bodies of existing data derived from tests of varied quality, but such data sets do not exist in the AUC format. Since obtaining sufficient data for adequate item calibration is anticipated to be a problem even for the Rasch model, it appeared sensible to concentrate initial efforts in this direction.

Model (i): Fixed Partial Credit

$$\text{The model is } E(X_{vi}) = \frac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$$

where: $E(X_{vi})$ is the expected score of person v on item i

α_v is a parameter describing the ability of person v

δ_i is a parameter describing the difficulty of item i

and the scoring function $X_{vi} = \frac{m_i - g_{vi}}{m_i - 1}$

where m_i is the number of alternative choices on item i (of which 1 is correct and (m-1) are incorrect)

and g_{vi} is the number of attempts by person v on item i until the correct alternative is chosen. If the (m_i-1)th attempt fails then $X_{vi}=0$.

The rationale for this scoring scheme is based on a "partial knowledge" distractor elimination model. If a correct response is chosen at the first attempt, then it is assumed that the student was able to eliminate all the distractors, and so he or she gets full credit. If the first attempt fails, but the second attempt is correct, it is assumed that he or she could eliminate all the distractors but one, so that credit of $\frac{m-2}{m-1}$ is awarded. (The number of distractors is (m-1)).

Although this equal-interval scoring function may appear somewhat arbitrary it is analogous to that frequently adopted in elementary scaling techniques (e.g., Likert scales). Moreover, Andersen (1977) has shown that for the model to retain specific objectivity, successive scoring categories must be equidistant. The immediate advantage of this is that the "raw score" by a student who has worked through the set of items is a sufficient statistic for the ability (and frequently may be used instead of it--hence the viability of the scheme proposed by Brown).

Parameter estimation is approached via a modification of the Rasch PAIR estimation algorithm (Choppin, 1982). For two items i and j, the relative difficulty can be estimated by

$$\delta_i - \delta_j = \log b_{ji} - \log b_{ij}$$

where, on this occasion, b_{ij} is the sum over all people in the sample, of $X_i(1-X_j)$ and b_{ji} is similarly defined. (It can be seen that this reduces to the standard PAIR algorithm in the case of 1/0 scoring.)

$X_i(1-X_j)$ represents the product of an estimate of the extent to which item i is mastered multiplied by an estimate of the extent to which item j is not mastered. It may be viewed, for each subject as a measure of the extent to which item i is easier than item j. The ratio:

$$\frac{E\left[X_i(1-X_j)\right]}{E\left[X_j(1-X_i)\right]} = e^{(\delta_j-\delta_i)} \qquad \text{a value independent of } \alpha$$

which is why the accumulation of data over persons to estimate these expectations works.

The algebra for maximum likelihood estimation, and for controlling the model via the squared matrix B* exactly duplicates that laid out in Choppin (1982), except that the formulae presented there for the standard errors of the δ-values are no longer appropriate. (Corrected formulae have not yet been developed, so the values reported by PAIR are used as conservative guides.) Once the items are calibrated, the estimation of person ability again follows the PAIR procedure.

Model (ii): Step Calibration

In this model, the probability of person v responding correctly to item i at the gth attempt, given that he or she makes the attempt, is:

$$\text{Prob.}\left[X_{vig} = 1\right] = \frac{e^{(\alpha_v - \delta_{ig})}}{1 + e^{(\alpha_v - \delta_{ig})}}$$

where X_{vig} = 1 if the gth attempt at item i is successful, and
0 otherwise

α_v is again a parameter describing the ability of person v and δ_{ig} is a parameter describing the difficulty of the gth step on item i.

For a five-way multiple choice item there are five possible sets of observation vectors $\underline{X}$, with asterisks indicating missing data (i.e., attempts that do not occur).

	g =	1	2	3	4
Correct at first attempt:	$\underline{X}$ =	1	*	*	*
Correct at second attempt:	$\underline{X}$ =	0	1	*	*
Correct at third attempt:	$\underline{X}$ =	0	0	1	*
Correct at fourth attempt:	$\underline{X}$ =	0	0	0	1
Failure at fourth attempt:	$\underline{X}$ =	0	0	0	0

If the raw data to be analyzed consists of code numbers for the successful attempt on each item, then it must be transformed into the above format for the calibration analysis. For example, suppose that an individual required (2, 1, 1, 4, 5, 3) attempts to find the correct answers to a six item five-way multiple choice test. The recoding of this vector would yield:

0 1 * *	1 * * *	1 * * *	0 0 0 1	0 0 0 0	0 0 1 *

a vector of 24 elements. A set of such vectors from the different persons attempting the test can be analyzed almost as a standard Rasch model problem--providing the PAIR algorithm (Choppin, 1982) is used to allow for the embedded missing data. The deviation from the standard Rasch procedure is necessitated by the violation of the local independence assumption for AUC data. While it remains important that between items this independence is maintained, it is clear that within an item the different X-values cannot be independent. As shown above, only m possible patterns out of the 3^m theoretically possible on each item ever occur and certain combinations such as (1,0) are impossible.

This invalidates the maximum likelihood estimation procedure which assumes that the elements of the B matrix for item pairs are essentially independent.

The full theoretical implications of this are still being explored, but a convenient "fix" in order to calibrate the items is to use instead of ML a least squares procedure based on a modified B* matrix. This B*, instead of being simply the square of matrix B as before, is now screened to remove the contaminating dependence within items.

In the standard PAIR algorithm

$$b^*_{ij} = \sum^{k} b_{ik}b_{jk}$$

and since $b_{ii} = b_{jj} = 0$, b^*_{ij} is independent of b_{ij}.

In PAIR as modified for AUC tests

$$b^*_{ij} = \sum^{k} v_{ik}b_{ik}v_{kj}b_{kj}$$

where v_{ik} are the elements of a screening matrix such that

$v_{pq} = 0$ if responses p and q relate to the same item

and $v_{pq} = 1$ otherwise.

Least squares estimation procedure applied to the B* matrix yields calibrations for the δ - values ($i = 1, k; g = 1, m-1$).

The estimation of person ability, the usual goal in such exercises, is somewhat different than in the standard Rasch model. Apart from rare failures at the final attempt, each student will score one point on each item and thus will have a raw score of k.

However, this raw score will be based on different numbers of "attempts", and individual step difficulties will be higher on some items than on others. Therefore α_v is estimated by the solution of

$$r_v - \sum \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_{ig}}} = 0$$

where the summation extends over the item steps actually attempted, and r_v is the observed raw score (usually k). This equation can always be solved to produce a unique LS estimation of α_v, but may be inefficient since its (iterative) solution is required for each observed score pattern. Monte Carlo simulation could compare the variation in α with the scoring function proposed by Brown (1965), to see whether the exact iterative solution is worthwhile.

The standard errors of such estimates depend upon the number of attempts made. Thus someone who usually responds correctly at the first attempt will be measured with less precision than someone who typically requires two or three attempts. Data in which the mean number of attempts per item is 2.0 (a typical value) will yield

standard errors of measurement only 0.7 times as large as with a conventional test with the same number of items. From this it can be seen that major increases in precision can only be achieved by substantially increasing the number of alternatives per question, so that the number of attempts made before success will also increase.

Model (iii): Distractor Calibration

This model is an extension of (ii) to allow for differences among the distractors. The item step difficulty parameter now describes the difficulty of the item at each step taking account of which distractors have already been eliminated.

Thus δ_{i1} indicates the difficulty of item i at the initial step when all distractors are present

$\delta_{i2.A}$ indicates the difficulty of item i at the second step when distractor A was chosen at the first

$\delta_{i3.BC}$ indicates the difficulty of item i at the third step after distractors B and C have been chosen (in whatever order)

With this notation, the model becomes

$$\text{Prob}\left[X_{vig.\underline{F}} = 1\right] = \frac{e^{(\alpha_v \delta_{ig.\underline{F}})}}{1 + e^{(\alpha_v - \delta_{ig.\underline{F}})}}$$

The analysis and estimation procedures essentially follow those for model (ii) except that the response data must be coded in

different format. For a five-way item (for which the correct response is E, and the distractors are labeled A-D), the structure of the parameters to be estimated is:

δ_{i1}	$\delta_{i2.A}$	$\delta_{i2.B}$	$\delta_{i2.C}$	$\delta_{i2.D}$	$\delta_{i3.AB}$	$\delta_{i3.AC}$	$\delta_{i3.AD}$	$\delta_{i3.BC}$	$\delta_{i3.BD}$	$\delta_{i3.CD}$	$\delta_{i4.ABC}$	$\delta_{i4.ABD}$	$\delta_{i4.ACD}$	$\delta_{i4.BCD}$

Response data for an individual who chose responses A, C, E, in that order, getting the item right at the third attempt, would be coded

0	0 * * *	* 1 * * * *	* * * *

It should be noted that this coding scheme is severely constrained. There is at most one entry in each block, and a "1" entry effectively terminates the vector. Thus the range of possible response patterns is limited, and again the local independence principle is violated.

Estimation procedures can follow the sequence described in model (ii) first to calibrate the item step values, and secondly to estimate the person ability parameters. However, it is apparent that the procedure is somewhat unwieldy. For each item the number of difficulty parameters to be estimated is given by $(2^{m-1} - 1)$ where m is the number of alternative responses in the item format. Inadequate calibration of the parameters due to insufficient data can spoil the overall measurement of person ability (viz: person measurement with

the Lord-Birnbaum three-parameter model and small data sets). A six item five-way multiple choice test such as that described under model (ii) would require the estimation of 90 item difficulty parameters under model (iii) as opposed to 24 under model (ii). For this model, in contrast to model (ii), it would seem wise to restrict item formats to not more than three or four alternatives.

3. Trial Data Analysis

Calibration procedures for models (i) and (ii) have been programmed in FORTRAN using variations of the PAIR algorithm described above. Both programs have demonstrated their ability to recover the parameter values used to generate artificial "fitting" data. Two data sets from AUC tests each comprising several hundred cases have been analyzed using these programs. One test is a junior high school science test under development in England. The second is a college level psychology test used in a private California university. The results are still being studied.

Model (iii) requires the coding of which distractors were selected in which sequence, and this is only practicable with a clinically administered or computer administered test. For this reason we have devoted considerable time to developing a software package that will administer AUC tests in schools, and store the results in a format suitable for aggregation and subsequent analysis. Details of this package are given in the Appendix.

REFERENCES

Andersen, E.B. Sufficient statistics and latent trait models. Psychometrika, 1977, 42, 69-81.

Brown, J. Multiple response evaluation of discrimination. The British Journal of Mathematical and Statistical Psychology, 1965, 18, 125-137.

Choppin, B.H. A fully conditional estimation procedure for Rasch model parameters. CSE Technical Report 196, Center for the Study of Evaluation, UCLA, 1983.

Frary, R.B. The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Psychological Measurement, 1980, 4, 1, 79-90.

Hanna, G.S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Ecu tional Measurement, 1975, 12, 3, 175-178.

Kane, M., & Moloney, J. The effect of guessing on item reliability under answer-until-correct scoring. Applied Psychological Measurement, 1978, 2, 1, 41-49.

Merwin, J.G. Rational and mathematical relationships of six scoring procedures applicable to three-choice items. Journal of Educational Psychology, 1959, 50, 4, 153-160.

Wilcox, R.R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74.

Wood, R. Multiple choice: A state of the art report. Evaluation in Education, 1977, 1, 191-280.